11 Survival Analysis and Empirical Likelihood

The first paper of empirical likelihood is actually about confidence intervals with the Kaplan-Meier estimator (Thomas and Grunkmeier 1979), i.e. deals with right censored data. Owen recognized the generality of the method and proved it rigorously for the non-censored data (1989). He also coined the term "Empirical Likelihood", and apply it to many other cases. Since then, many researches have been done and establish the empirical likelihood as a general method with many nice features. Most of the research is about non-censored data, only a small portion is about the survival analysis.

See Owen's Book of 2001.

We will mainly discuss the empirical likelihood in the context of survival analysis here.

11.1 What is the empirical likelihood method?

It is the non-parametric counter part of the Wilks likelihood ratio approach in the parametric likelihood inference. (if you are familiar with that). You do not need to pick a parametric family of distributions to do the inference. The primary reference is Owen's Book of 2001 *Empirical Likelihood*.

It helps you to do statistical inference (calculate P-value/confidence intervals) with certain non-parametric tests/estimators without the need of calculating their variances. In this respect, it is similar to Bootstrap method.

11.2 Why empirical likelihood with survival analysis?

Because the variance is so much harder to estimate for NPMLE with censored/truncated data. Because empirical likelihood method do not need a variance estimator. Because empirical likelihood inference is often (at least asymptotically) efficient. For specific examples demonstrate these points, see sections later.

11.3 The empirical likelihood function for censored data

The empirical likelihood method is intrinsically linked to the NPMLE. We first study the empirical likelihood function for censored data.

Definition of Empirical Likelihood, The Nelson-Aalen, Kaplan-Meier estimator is NPMLE:

Let $(x_1, \delta_1), (x_2, \delta_2), \ldots, (x_n, \delta_n)$ be the i.i.d. right censored observations as defined in class: $x_i = \min(T_i, C_i), \delta_i = I[T_i \leq C_i]$, where T_i are lifetimes and C_i are censoring times.

Let the CDF of T_i be $F(t)$.

The likelihood pertaining to $F(\cdot)$ based on the right censored samples (x_j, δ_j) is (constant has been omitted)

$$L(F) = \prod_{\delta_i=1} \Delta F(x_i) \prod_{\delta_i=0} (1 - F(x_i)).$$

Because the survival function $(1 - F(t))$ and hazard function $(\Lambda(t))$ are mathematically equivalent, inference for one can be obtained though a transformation of the other.

For discrete CDF, we have

$$\begin{cases} \Delta F(x_i) & = \Delta\Lambda(x_i) \prod_{j:x_i<x_j} (1 - \Delta\Lambda(x_j)) \\ 1 - F(x_i) & = \prod_{j:x_j \le x_i} (1 - \Delta\Lambda(x_j)) \end{cases},$$

therefore, while assuming F is discrete,

$$L(\Lambda) = \prod_{i=1}^{n} \left\{ (\Delta\Lambda(x_i))^{\delta_i} \Big(\prod_{j:x_j<x_i} (1 - \Delta\Lambda(x_j)) \Big)^{\delta_i} \Big(\prod_{j:x_j\le x_i} (1 - \Delta\Lambda(x_j)) \Big)^{1-\delta_i} \right\} .$$

Let $x_{k,n}$ be the m distinctive and ordered values of the x-sample above, where $k = 1 \cdots m$ and $m \le n$. Let us denote

$$D_k = \sum_{i:x_i=x_{k,n}} \delta_i \qquad \text{and} \qquad Y_k = \sum_{i=1}^{n} I(x_i \ge x_{k,n}).$$

Then, L becomes

$$L = \prod_{k=1}^{m} (\Delta\Lambda(x_{k,n}))^{D_k} (1 - \Delta\Lambda(x_{k,n}))^{(Y_k - D_k)}. \tag{10}$$

From this, we can easily get the log likelihood function:

$$\log L = \sum D_k \log \Delta\Lambda(x_{k,n}) + (Y_k - D_k) \log(1 - \Delta\Lambda(x_{k,n})) .$$

Maximize the log likelihood function wrt $\Delta\Lambda(x_{k,n})$ gives the (nonparametric) maximum likelihood estimate of $\Delta\Lambda(x_{k,n})$ which is the Nelson Aalen estimator,

$$\Delta\hat{\Lambda}(x_{k,n}) = \frac{D_k}{Y_k}. \tag{11}$$

By the invariance property of the MLE this implies that the Kaplan-Meier estimator

$$1 - \hat{F}(t) = \prod_{x_{k,n}\le t} (1 - \frac{D_k}{Y_k})$$

is the MLE of $1 - F(t) = \prod_{s\le t}(1 - \Delta\Lambda(s))$ (i.e. it maximizes the likelihood $L(F)$ above.)

If we use a 'wrong' formula connecting the CDF and cumulative hazard: $1 - F(t) = \exp(-\Lambda(t))$, (wrong: in the sense that the formula works for continuous F but here we have a discrete F) then the likelihood would be

$$AL = \prod_{i=1}^{n} (\Delta\Lambda(x_i))^{\delta_i} \exp\{-\Lambda(x_i)\} = \prod_{i=1}^{n} (\Delta\Lambda(x_i))^{\delta_i} \exp\{- \sum_{j:x_j\le x_i} \Delta\Lambda(x_j)\}.$$

If we let $w_i = \Delta\Lambda(x_i)$ then

$$\log AL = \sum_{i=1}^{n} \delta_i \log w_i - \sum_{i=1}^{n} \sum_{j: x_j \le x_i} w_j \ . \tag{12}$$

It is worth noting that among all the cumulative hazard functions, the Nelson-Aalen estimator also maximizes the $\log AL$,

$$w_i = w_i^* = \frac{D_i}{Y_i} \ . \tag{13}$$

This can be verified easily by taking derivative of $\log AL$ wrt w_i and solving the equation.

This version of the likelihood is sometimes called the 'Poisson' version of the empirical likelihood.

11.4 The property of the NPMLE. Computation. Self-consistent algorithm.

Many of the properties of NPMLE can be obtained by using the tools of counting process martingales and empirical process theory.

In simple cases (like in the previous section) the NPMLE has explicit formula. In more complicated cases (more complicated censoring, or with some extra constraints) no explicit formula exist and the computation of the NPMLE can be done by self-consistency/EM algorithm. See Zhou 2002 for details.

In particular, using the counting process/martingale theory, we can show that

Lemma 1 *Let $\hat\Lambda(t)$ denote the Nelson-Aalen estimator. Under the conditions that makes the Nelson-Aalen estimator $\hat\Lambda(t) - \Lambda(t)$ a martingale, we have*

$$\sqrt{n} \int_0^\infty g_n(t) d[\hat\Lambda(t) - \Lambda(t)] \xrightarrow{\mathcal{D}} N(0, \sigma^2) \tag{14}$$

where $g_n(t)$ is a sequence of $\mathcal{F}_t$ predictable functions that converge (in probability) to a (non-random) function $g(t)$, provided the variance at right hand side is finite.

The variance σ^2 above is given by

$$\sigma^2 = \int_0^\infty \frac{g^2(s) d\Lambda(s)}{(1 - F(s-))(1 - G)} \tag{15}$$

Also, the variance σ^2 can be consistently estimated by

$$\sum_{i=1}^{n} n \left[\frac{g_n^2(x_i)}{Y_i} \frac{D_i}{Y_i} \right] \ . \tag{16}$$

Proof: If we put a variable upper limit, v, of the integration in (14), then it is a martingale in v. By the martingale CLT, we can check the required conditions and obtain the convergence to normal limit at any fixed time v. Finally let v equal to infinity or a large value.

To get the variance estimator, we view the sum as a process (evaluated at ∞) and then compute its compensator, which is also the predictable variation process. Then the compensator or intensity is the needed variance estimator. Finally, use Lenglart inequality to show the convergence of the variance estimator. $\diamondsuit$

11.5 Maximization of the empirical likelihood under a constraint

We now consider maximization of the empirical likelihood function above for right censored observations with an extra estimating equation constraint.

We will maximize the $\log AL$ among all cumulative hazard functions that satisfy

$$\int g(t) d\Lambda(t) = \sum g(x_j) \Delta\Lambda(x_j) = \mu \ . \tag{17}$$

Detailed proof is only given in this case here.

We may also maximize the $\log L$ among all cumulative hazard functions that satisfy

$$\sum g(x_j) \log(1 - \Delta\Lambda(x_j)) = \mu \ .$$

For detailed proof of this case, see Fang (2000). (Add some comments/examples about this equation.)

Theorem 11.1 *The maximization of the log likelihood* $\log AL$ *under the extra equation (17) is achieved when*

$$\Delta\Lambda(x_k) = \tilde{w}_k = \frac{D_k}{Y_k + \lambda g(x_k)} \tag{18}$$

with λ *obtained as the solution of the equation*

$$\sum_{k=1}^{n} g(x_k) \frac{D_k}{Y_k + \lambda g(x_k)} = \mu \ . \tag{19}$$

Proof: Use the Lagrange multiplier to compute the constrained maximum. Notice the equation for λ is monotone so there is a unique solution. $\diamondsuit$

Sometimes we may want to use the following estimating equation instead:

$$\sum g(x_j) \Delta F(x_j) = \mu \ .$$

i.e. maximize the $\log L(F)$ among all CDF that satisfy this (mean type) equation. Similar results also hold, but harder to proof. Computation is also harder in this case. No explicit formulae exist for the $\Delta F(x_j)$ that achieve the max under the constraint equation (as far as I know). We have to use a modified self-consistent/EM algorithm to compute the NPMLE under this constraint, available in the R package `emplik`.

11.6 Wilks Theorem For Censored Empirical Likelihood

The (ordinary) Wilks theorem in parametric estimation says:

$$-2\log \frac{\max\limits_{\theta\in\Theta_0} L(\theta)}{\max\limits_{\theta\in\Theta} L(\theta)} \approx \chi^2_p$$

where Θ is the whole parameter space, and Θ_0 is the subspace of Θ, obtained by imposing p equations for θ to satisfy.

The right-hand side is the usual chi-square if the true θ is inside Θ_0 (i.e. when the null hypothesis is true), and is a non-central chi-square when the true θ is not inside Θ_0.

The likelihood function $L(\theta)$ above is the ordinary (parametric) likelihood function.

There is a similar Wilks theorem where the likelihood function above is replaced by the (censored) empirical likelihood function. The whole parameter space Θ will be equal to "all the CDFs" or "all the cumulative hazard functions". and the subspace Θ_0 is obtained by imposing equations like (17) or (20) on Θ.

We only prove a Wilks theorem for the censored empirical likelihood $\log AL$ and estimation equation (17).

Theorem 11.2 *Assume we have iid right censored observations as before.*

Then, as $n \to \infty$

$$-2\log \frac{AL(w_i = \tilde{w}_i)}{AL(w_i = w_i^*)} \xrightarrow{\mathcal{D}} \chi^2_1$$

where w_i^* is the Nelson Aalen estimator, and $\tilde{w}_i$ is given by (18) with $\mu = \mu_0 = \int g(t)d\Lambda_0$.

We first prove a lemma.

Lemma 11.2 *The solution λ^* for the equation (19) above has the following asymptotic distribution.*

$$\frac{\lambda^*}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, var =)$$

PROOF: (Basically, we do a one step Taylor expansion.) Assume λ is small (which can be shown), we expand (around $\lambda = 0$) the left hand side of the equation (19) that defines λ^*, we can re-write the equation as

$$\sum \frac{g(x_k)D_k}{Y_k} - \lambda \sum \frac{g^2(x_k)D_k}{(Y_k)^2} + O(\lambda^2) = \mu$$

Ignor the higher order term, this equation can be solved easily, we therefore have:

$$\lambda^* = \frac{n\sum_k \frac{g(x_k)D_k}{Y_k} - \mu}{n\sum \frac{g^2(x_k)D_k}{(Y_k)^2}} + O_p(\lambda^2)\ .$$

Multiply $1/\sqrt{n}$ throughout, we notice that the numerator is same as the one treated in Lemma 1, and thus have an asymptotic normal distribution for numerator. The denominator is seen to converge in probability to σ^2 also by Lemma 1. Thus by Slutsky theorem we have the conclusion. $\diamondsuit$

Now we are ready to proof theorem 11.2.

PROOF OF THEOREM 11.2: Plug the expressions of $\tilde{w}_i$ and w_i^* into $AL(w_i)$ or $\log AL$.

$$-2\log\frac{AL}{AL} = 2[\log AL(\lambda=0) - \log AL(\lambda=\lambda^*)]$$

Now we use Taylor expansion (up to second order) on the second term above: expand at $\lambda = 0$.

The first term in the expansion will cancel with the other term.

The second term in the expansion is zero. Because the (first order) partial dervative wrt λ at $\lambda = 0$ is zero. This is also obvious, since $\lambda = 0$ gives rise to the Nelson-Aalen estimator, and the Nelson-Aalen estimator is the maximizer. The derivative at the maximizer must be zero.

The third term in the expansion is equal to

$$-\frac{\partial}{\partial\lambda^2}\cdot(\lambda^*)^2\sum\frac{D_k g^2(x_k)}{Y_k^2}$$

which can be shown to be approximately chi-square by Lemma 2 and the variance estimator. (plug in the asymptotic representation of λ^*).

We see that the limiting distribution is chi-square. $\diamondsuit$

Multivariate version of this theorem exists. Zhou (2002)

This theorem can be used to test the hypothesis $H_0 : \int g(t)d\Lambda(t) = \mu$. We reject the null hypothesis when the $-2\log$ empirical likelihood ratio is too large.

11.7 Confidence interval/region by using Wilks theorem

The above section points out a way to do testing hypothesis. To obtain confidence intervals we need to invert the testing.

The software we use in the example is a user contributed package `emplik` of R. They can be obtained from http://www.cran-us.org

Obtain confidence interval for one of the two parameters.

Example: See my notes "Survival Analysis using R".

11.8 Empirical likelihood for regression models (Cox model)

11.9 AFT models

11.10 Referrences

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, **75** 237-249.

Owen, A. (1990), "Empirical Likelihood Confidence Regions," *The Annals of Statistics*, **18**, 90-120.

Owen, A. (2001). *Empirical Likelihood.* Chapman and Hall, New York.

Pan, X. and Zhou, M. (2002). "Empirical likelihood ratio in terms of cumulative hazard function for censored data" *Journal of Multivariate Analysis.* 166-188.